# Accounting for Chance Agreement in Gesture Elicitation Studies

Theophanis Tsandilas, Pierre Dragicevic

# Accounting for Chance Agreement in Gesture Elicitation Studies

**Theophanis Tsandilas**[1,2,3] **Pierre Dragicevic**[1,3]
theophanis.tsandilas@inria.fr pierre.dragicevic@inria.fr
[1]Inria [2]Univ Paris-Sud, CNRS (LRI) [3]Université Paris-Saclay

## ABSTRACT
The level of agreement among participants is a key aspect of gesture elicitation studies, and it is typically quantified by means of agreement rates (AR). We show that this measure is problematic, as it does not account for chance agreement. The problem of chance agreement has been extensively discussed in a range of scientific fields in the context of inter-rater reliability studies. We review chance-corrected agreement coefficients that are routinely used in inter-reliability studies and show how to apply them to gesture elicitation studies. We also discuss how to compute interval estimates for these coefficients and how to use them for statistical inference.

### Author Keywords
Gesture elicitation, agreement rate, agreement coefficient, chance agreement, kappa coefficients, confidence intervals

### ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (HCI)

## INTRODUCTION
Gesture elicitation is widely used in HCI as a method for designing better gesture-based interfaces by involving end users [21]. This method seeks to identify gesture vocabularies that are self-discoverable or easy to learn. Participants are typically shown the outcome of user interface actions or commands, and asked to propose gestures that would trigger these actions. The hope is that consistent gesture/command associations will emerge, but participants can also be very inconsistent in their proposals. Thus analyzing *agreement* between participants is a key aspect of the method [21, 18]. Agreement can help assess whether the use of gestures is appropriate for a certain situation, can guide the design and evaluation of gesture vocabularies, and can help understand why some commands naturally map to gestures while others do not.

The most widely used index for quantifying agreement in gesture elicitation studies is Wobbrock et al's coefficient $A$ [20], recently superseded by Vatavu and Wobbrock's *agreement rate* $AR$ [18]. The authors convincingly argue for the use of $AR$ rather than $A$, derive a significance test for making comparisons between $AR$ values, and suggest conventional ranges of values (low, medium, high) to help researchers interpret levels of agreement and compare them across studies.

However, it has long been recognized in other disciplines that $AR$ is problematic because it does not account for chance agreement [3, 7, 9]. In this article we explain the issue of chance agreement and how it can affect $AR$. We review previous work on inter-rater reliability assessment, where these issues have been extensively studied. Several chance-corrected agreement coefficients have been proposed [9] and are routinely used in a range of areas including psychometrics, medical research, computational linguistics, as well as in HCI for video and user log analysis [10]. We discuss how to use chance-corrected agreement coefficients in gesture elicitation studies, how to compute and interpret interval estimates, and conclude with pending challenges and limitations.

## AGREEMENT MEASURES IN ELICITATION STUDIES
Wobbrock et al. [21] refer to presented commands as *referents*, and assume that all user-elicited *gestures* can be categorized into classes of equivalence they call *signs*. To quantify agreement for a given referent $i$, a great number of elicitation studies have used the formula initially proposed by Wobbrock et al. [20] in their early 2005 paper on gesture elicitation:

$$A_i = \sum_{k=1}^{q} \left( \frac{n_{ik}}{n_i} \right)^2 \quad (1)$$

where $q$ is the total number of signs produced in the study, $n_{ik}$ is the number of participants who proposed sign $k$ for the referent $i$, and $n_i$ is the total number of gestures proposed for the referent $i$. In the common situation where all participants are presented with all referents, $n_i$ is the number of participants in the study. To obtain the overall agreement $A$, Wobbrock et al. [20] then average $A_i$ across all referents.

Later on, Vatavu and Wobbrock refine this index of agreement $A_i$ and propose to replace it with a slightly different index $AR_i$ they call the *agreement rate*, defined as follows:

$$AR_i = \sum_{k=1}^{q} \frac{n_{ik}(n_{ik}-1)}{n_i(n_i-1)} \quad (2)$$

Vatavu and Wobbrock point out that in contrast to the $A_i$ index, $AR_i$ takes values in the entire interval $[0..1]$ and has a clear interpretation whereby $AR_i$ is the proportion of participant pairs who are in agreement. It can be further observed that $AR_i$ reduces to $A_i$ for large samples. Both $AR_i$ and its approximation $A_i$ have been used in a range of disciplines as an index of homogeneity for nominal data, and are most commonly referred to as the *Simpson's index* [14, 6].

The content of this paper was finalized on September 22, 2015. It was registered as research report on February 2016: Research Report No 1584, Laboratoire de Recherche en Informatique (LRI), CNRS and Université Paris-Sud, Feb 2016.

| | R1 | R2 | R3 | R4 | R5 | R6 | R7 | R8 | R9 | R10 | total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **A** | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 5 |
| **B** | 7 | 2 | 6 | 4 | 10 | 3 | 10 | 3 | 10 | 5 | 60 |
| **C** | 6 | 5 | 9 | 4 | 5 | 10 | 3 | 10 | 3 | 5 | 60 |
| **D** | 4 | 6 | 2 | 4 | 4 | 4 | 3 | 4 | 4 | 5 | 40 |
| **E** | 3 | 6 | 3 | 7 | 0 | 2 | 4 | 3 | 3 | 4 | 35 |

**Table 1. Table summarizing the results of a fictitious gesture elicitation study, where 20 participants propose a grasp (A, B, C, D, E) for 10 different referents (R1 to R10). Each cell shows the number of participants who chose the given grasp for the given referent.**

As before, Vatavu and Wobbrock average $AR_i$ across all referents to obtain an *overall agreement rate* $AR$. This index also has a long history in science, and is commonly referred to as the *percent agreement* [9]. However, it is also widely known to be problematic [7, 9], as we will now illustrate.

## THE PROBLEM OF CHANCE AGREEMENT

Suppose a researcher is interested in a new ten-inch tablet that can sense user grasps, and wants to see if previously proposed techniques [19] can be improved through gesture elicitation. She recruits 20 participants to whom she shows 10 different document navigation operations in the form of animations on the tablet, and asks them to choose among five different grasps [19]. Table 1 summarizes the results.

A visual inspection of this table reveals a mix of disagreement and agreement. The researcher is initially worried, but is then relieved to see that her agreement analysis yields a quite respectable $AR = .265$. This value can be considered respectable because it is slightly higher than the average $AR$ reported by Vatavu and Wobbrock [18] from 18 previous elicitation studies, and because it can be interpreted as a *medium* level of agreement according to their recommendations.

The researcher was right to be worried, however. Suppose that the study is replicated, but participants are blindfolded and cannot see any of the referents presented to them — they are simply asked to guess. Their choice of grasp will thus be random. Suppose all grasps are equally preferred. Shockingly, this study will yield an expected overall agreement rate of $AR = .200$, not far from the previously observed $AR$.

Since all grasps are equally likely, the probability that a given participant proposes a specific grasp for a given referent is $1/5 = 0.2$, and the probability that a given pair of participants proposes it is $0.2 \times 0.2$. Since two participants can agree on any of the five grasps, the probability of agreement for a pair of participants on a referent is $5 \times 0.2 \times 0.2 = 0.2$. Therefore, the expected proportion of participant pairs who are in agreement — that is, the expected $AR$ — is 0.2.

Given that there could not have been any intrinsic agreement between participants, one would rather expect an agreement index to give a result close to zero. Furthermore, one would certainly not label such a result as a "medium" agreement.

Arguably, the blindfolded study is purely fictional, and no gesture elicitation study involves participants who make completely random decisions. Nevertheless, gesture elicitation involves subjective judgments, where randomness can play a role. A participant may be uncertain about which gesture is the best, and in some situations, may even respond randomly. Such situations can include highly abstract referents for which there is no intuitive gesture, poor experimental instructions, gesture options that are too similar, or a lack of user familiarity with the specific domain or context of use.

Due to sources of randomness in participants' choice of signs, any value of $AR$ reflects both intrinsic and spurious agreement. How much spurious agreement $AR$ reflect depends on the likelihood of chance agreement, which in turn depends on the number of signs. The vocabulary of five signs used in our example is rather small, but not implausible for a study. Even for large vocabularies, participants often show a strong preference for a few signs (e.g., [1]). Such biases can greatly increase chance agreement and inflate agreement rates. Therefore, it seems safer to always correct for chance agreement.

## CORRECTING FOR CHANCE AGREEMENT

A large volume of research has examined the problem of chance agreement in the context of inter-rater reliability studies, i.e., studies that involve subjective human assessments [9]. In this section, we explain how results from this research can be applied to gesture elicitation studies.

Inter-rater reliability studies employ a different terminology from gesture elicitation studies, but the mapping between the two is straightforward. Study participants become *raters* (also called judges or coders), referents become *items* (often confusingly called subjects), signs become *categories*, and gestures become *ratings* (or judgments) [9].

### Kappa Coefficients

Work on chance-corrected indices of agreement dates back to at least the 50–60's. Early on, Cohen proposed the Kappa coefficient to measure the agreement between two raters:

$$\kappa = \frac{p_a - p_e}{1 - p_e} \tag{3}$$

where $p_a$ is the proportion of items on which both raters agree, and $p_e$ is chance agreement. According to Cohen [3], Kappa measures *"the proportion of agreement after chance agreement is removed from consideration."* The nominator captures *"the percent of units in which beyond-chance agreement occurred"*, while the denominator is a normalizing term that captures maximum beyond-chance agreement.

The $p_e$ term does not assume equiprobable categories as in our blindfolded study but instead considers the distribution of categories assigned by each rater across all items. Scott's $\pi$ coefficent [17] makes the additional assumption that all raters share the same distribution of categories.

Most chance-corrected agreement indices known today are based on Equation 3. Each index makes different assumptions and has different limitations [9]. More than forty years ago, Fleiss [7] proposed to extend Cohen's Kappa (but actually extended Scott's $\pi$) to multiple raters. For $p_a$ he uses the *"proportion of agreeing pairs out of all the possible pairs of assignments"* [7], also called *percent agreement*. This formulation for $p_a$ is used in many other indices, and is the same as Vatavu and Wobbrock's [18] *agreement rate* $AR$.

For the chance agreement term $p_e$, Fleiss uses:

$$p_e = \sum_{k=1}^{q} \pi_k^2, \qquad \pi_k = \frac{1}{m} \sum_{i=1}^{m} \frac{n_{ik}}{n_i} \tag{4}$$

where $m$ is the total number of items, $n_{ik}$ is the number of ratings for item $i$ having category $k$, and $n_i$ is the total number of ratings for item $i$. The value $\pi_k$ estimates the probability that a rater classifies an item into category $k$, based on how many times this category has been used across the entire study.

In Table 1, Fleiss' chance agreement is $p_e = .251$. Fleiss' Kappa is thus $\kappa_F = \frac{.265-.251}{1-.251} = .018$, reflecting a close-to-chance overall agreement. Note that $\kappa_F$ is allowed to be negative, which would suggest a beyond-chance disagreement.

A large volume of research has been devoted on how to best estimate $p_e$. The Brennan-Prediger coefficient [2] uses the simplest estimate $p_e = 1/q$, where $q$ is the total number of categories. It assumes equiprobable selection of categories, as in our blindfolded example. The chance agreement for Table 1 becomes $p_e = .200$, and the Brennan-Prediger agreement coefficient yields $\kappa_q = .081$. However, Table 1 suggests that equiprobable categories is not a realistic assumption, as in most gesture elicitation studies. The coefficient has also been criticized for giving researchers incentive to increase the number of categories to artificially inflate agreement.

Another measure of agrement, widely used in content analysis, is Krippendorff's $\alpha$ [13]. Krippendorff's $\alpha$ uses a different formulation for both $p_a$ and $p_e$ and can be used for studies with any number of raters, incomplete data (i.e., not all raters rate all items), and different scales including nominal, ordinal and ratio. For simple designs, its results are generally very close to Fleiss' Kappa, especially when there is no missing data and the number of raters is greater than five [9].

#### Applying Agreement Coefficients to Elicitation Studies

Any chance-corrected agreement coefficient can be used to analyze agreement in gesture elicitation studies. Gesture elicitation studies, however, have unique features. For example, in contrast to typical inter-rater reliability studies where raters choose among a pre-assigned set of categories, many gestures elicitation studies do not enforce a fixed set of categories. Participants can be creative and propose gestures that are not foreseen by the investigators. In these cases, categories are defined a posteriori by the investigators, after inspecting the data. In the simplest cases where gestures can still be categorized objectively, the validity of agreement rates should not be affected. More complex cases are outside the scope of this note and will be briefly discussed in the conclusion.

Research on inter-rater agreement has also mostly focused on the use of overall agreement scores for validation purposes. In contrast, gesture elicitation studies are mainly used to inform design. Researchers are interested in finer details concerning agreement, i.e., situations in which agreement is high and situations which exhibit little consensus. To this end, the analysis and graphing of agreement scores for all individual items (referents) is a useful and commonly employed method. When examining agreement on individual referents or groups of referents, chance-corrected agreement can be assessed by computing $p_a$ for each referent or group of referent, and using a common $p_e$ estimated across all items. Agreement indices also exist that are computed on an item-per-item basis [16].

Since chance correction typically only scales and offsets all per-referent agreements, relative differences are preserved. Thus, if only relative differences are of interest (e.g., which are the most consensual and the least consensual referents?), the use of percent agreements is acceptable. Correction for chance agreement is more strongly recommended when overall chance agreements are reported, or when comparing agreements across different gesture sets or different studies.

#### Statistical Inference

Statistical inference is the process of drawing conclusions about populations by observing random samples. When doing so, it is crucial to determine what is randomly sampled and what is not. In many inter-rater reliability studies, items are assumed to be sampled from a larger population of items [9]. In gesture elicitation studies, items (i.e., referents) are fixed: any conclusion typically only applies to these items. Raters, in contrast, are chosen arbitrarily, and an investigator may need to generalize her conclusions to the entire population of potential raters. Rater recruitment is a source of variability in the calculation of agreement coefficients, and this variability should ideally be acknowledged when presenting results, and when comparing results across studies.

The sampling distribution of agreement coefficients is often hard to approximate, but resampling methods such as jackknifing [9] and bootstrapping [11, 22] can be used to produce variance estimates, standard errors and confidence intervals for almost any agreement coefficient. Confidence intervals can be used both for communicating uncertainty and for testing hypotheses [5]. To compare two agreement scores obtained from the same set of participants (e.g., to compare agreement across referents or groups of referents), a jackknife or bootstrap confidence interval can be computed on their difference. This generally yields higher statistical precision than inspecting how individual confidence intervals overlap [4].

#### Interpreting the Magnitude of Agreement

Gwet [9] dedicates a full chapter on how to interpret the magnitude of an agreement. Several authors suggested conventional thresholds to help researchers in this task – Fleiss, for example, labels $\kappa < .400$ as "poor" and $\kappa > .750$ as "excellent" [9]. Krippendorff suggested $\alpha > .667$ and then later $\alpha > .800$ as thresholds below which data must be rejected as unreliable [12]. However, he and many others recognized that such thresholds are largely arbitrary and should be chosen depending on the application domain and on the *"costs of drawing invalid conclusions from these data"* [12]. It has also been emphasized that the magnitude of an agreement cannot be interpreted if confidence intervals are not provided [9, 12].

In gesture elicitation studies, the bar for an agreement score to be considered acceptable is way lower, even when ignoring chance agreement by considering $AR$ only. Vatavu and Wobbrock [18] offer guidelines for interpreting $AR$ based on a probabilistic reasoning and a survey of past studies.

| | A | AR | Fleiss' Kappa |
|---|---|---|---|
| **Keys** | .320 [.213, .427] | .284 [.172, .397] | .260 [.148, .371] |
| **Gestures** | .370 [.323, .417] | .336 [.287, .386] | .240 [.192, .289] |

**Table 2. Values of different indices of agreement ($A$, $AR$ and Fleiss' $\kappa$) for Bailly et al.'s study [1]. Brackets indicate 95% jackknife CIs.**

They suggest to refer to $AR < .100$ as a low agreement, $.100 < AR < .300$ as medium, $.300 < AR < .500$ as high, and to $AR > .500$ as a very high agreement.

However, Vatavu and Wobbrock's probabilistic reasoning is based on a null distribution assuming a 50% chance of agreement between participant pairs, so it is not clear to what extent it can answer a question about effect sizes. Setting standards based on results from past studies seems more sensible, but can also discourage efforts to raise our standards. Indeed, there does not seem to be any valid reason to be satisfied with a gesture agreement rate of .2 or .5. As much as we would like to have objective rules to help us distinguish between acceptable and unacceptable agreement scores, it is wise to refrain from using any such rule until these can be grounded in more solid cost-benefit analyses that integrate usability metrics.

## CASE STUDY

We demonstrate the use of chance-corrected agreement coefficientsby re-analyzing a gesture elicitation study by Bailly et al. [1]. The study was also re-analyzed by Vatavu and Wobbrock [18], thus it provides a good basis for comparison. Bailly et al introduce Métamorphe, a keyboard with actuated keys that sense user gestures (e.g., pull, twist, or push sideways). In their study, 20 users suggested a keyboard shortcut for 42 referents on a Métamorphe mockup. Choosing a keyboard shortcut required choosing *i)* a *key* and *ii)* the *gesture* applied to the key. Bailly et al. treat shortcuts as a whole, but also analyze keys and gestures separately. Here, we focus on analyzing keys and gestures separately. Participants produced a total of 71 different signs for keys and 27 different signs for gestures (compound gestures were counted as separate signs).

Table 2 shows overall agreement scores for keys and for gestures, computed using $A$, $AR$ and Fleiss' Kappa[1]. An investigator who uses the index $A$ or $AR$ may infer that participants' consensus was higher for gestures than for keys. However, Fleiss' Kappa values reveal that this difference is most likely due to chance agreement. As the number of signs was lower for gestures than for keys and participants exhibited a strong bias for the "top push" sign [1], agreement was more likely to occur by chance. Fleiss' Kappa corrects for this and allows the investigator to reason about intrinsic agreement.

In addition to point estimates of agreement values, Table 2 reports 95% CIs calculated using the jackknife method [9]. Researchers attached to null hypothesis significance testing can observe that none of the intervals contains zero, thus the existence of agreement (both chance-corrected and uncorrected) can be presented as statistically significant at the $\alpha$=.05 level [5]. However, showing that an agreement is greater than zero is not very informative [12, 9]. CIs allow for more useful statements, e.g., we can be reasonably confident that across all users similar to the ones recruited for the study, the average Fleiss' Kappa is between .15 and .37 for keys, and between .19 and .29 for gestures. We can further test whether the difference we noticed between the $AR$s for keys and gestures is reliable: the difference is 0.05, 95% CI = [-0.05, 0.16], thus we cannot confidently conclude that there is a difference. Finally, we can back up claims on the effect of referents such as *"highly directional commands [...] tended to have a high gesture agreement"* [1]. The difference in Fleiss' Kappa between the 8 referents containing the terms *top*, *bottom*, *left*, *right*, *previous* or *next* and all other referents is 0.41, 95% CI [0.24, 0.58], so the evidence is overwhelming.

[1]Bailly et al only considered 15 signs for *gestures* by grouping uncommon signs into "combo" and "other". We kept all 27 signs to be consistent with Vatavu and Wobbrock's analysis, but the results are very similar. Vatavu and Wobbrock report $AR = .336$ but $A = .406$ for gestures, which is likely an error since it violates the linear relationship between $A$ and $AR$ expressed in their Equation (3).

## DISCUSSION AND CONCLUSION

We discussed the problem of chance agreement in gesture elicitation studies, and how it can bias results. This issue most strongly affects studies with small sign vocabularies, and studies exhibiting a strong user bias for some signs. Since user bias cannot be controlled for, it is safer to always use indices that correct for chance agreement. We also motivated the use of confidence intervals, especially when making claims about overall agreement scores.

There are many issues this note does not address. For example, indices of agreement between user-elicited gestures and UI gesture sets such as the "guessability" metric proposed by Wobbrock et al. [20] are important but not addressed here. Furthermore, gesture elicitation studies can employ complex designs features that are not well supported by existing inter-rater agreement methods or may violate their model assumptions. These include the use of non-overlapping gestures sets and semantic grouping [1], hierarchically-structured sign sets [1], and multiple gesture proposals per referent [15].

Another pending issue is how to account for the subjectivity often inherent in the process of classifying user-elicited gestures into signs, especially in open coding settings [8]. This subjectivity poses a threat to validity and can possibly render all data and analyses meaningless, agreement indices included. Content analysis has developed methods to ensure reliability in similar situations [13], but these have been largely ignored in elicitation studies. Since content analysis also uses agreement indices, gesture elicitation studies would need to consider agreement at two very different levels.

Gesture elicitation studies are extremely useful but can be very complex to set up and analyze. The proper methodology that can ensure reliability and scientific rigor largely remains to be developed. HCI can gain a lot by considering the lessons learned in other disciplines where similar issues have been discussed, instead of (re-)developing methods in isolation. HCI also has the opportunity to contribute to the interdisciplinary debate on how to assess and use agreement, given that its complex study designs can pose interesting methodological challenges that have never been studied before.

## REFERENCES


1. Bailly, G., Pietrzak, T., Deber, J., and Wigdor, D. J. Métamorphe: Augmenting hotkey usage with actuated keys. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, ACM (New York, NY, USA, 2013), 563–572.
2. Brennan, R. L., and Prediger, D. J. Coefficient kappa: Some uses, misuses, and alternatives. *Educational and psychological measurement 41*, 3 (1981), 687–699.
3. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement 20*, 1 (1960), 37.
4. Cumming, G., and Finch, S. Inference by eye: confidence intervals and how to read pictures of data. *American Psychologist 60*, 2 (2005), 170.
5. Dragicevic, P. HCI Statistics without p-values. Tech. Rep. RR-8738, Inria, June 2015.
6. Ellerman, D. History of the logical entropy formula. Online, 2010. http://www.ellerman.org/history-of-the-logical-entropy-formula/.
7. Fleiss, J. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76*, 5 (1971), 378–382.
8. Grijincu, D., Nacenta, M. A., and Kristensson, P. O. User-defined interface gestures: dataset and analysis. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*, ACM (2014), 25–34.
9. Gwet, K. *Handbook of Inter-Rater Reliability, 4th Edition: The Definitive Guide to Measuring The Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.
10. Hailpern, J., Karahalios, K., Halle, J., Dethorne, L., and Coletto, M.-K. A3: Hci coding guideline for research using video annotation to assess behavior of nonverbal subjects with computer-based intervention. *ACM Transactions on Accessible Computing (TACCESS) 2*, 2 (2009), 8.
11. Hayes, A. F., and Krippendorff, K. Answering the call for a standard reliability measure for coding data. *Communication methods and measures 1*, 1 (2007), 77–89.
12. Krippendorff, K. Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research 30*, 3 (2004), 411–433.
13. Krippendorff, K. *Content analysis: An introduction to its methodology*. Thousand Oaks, CA: Sage, 2013.
14. Lieberson, S. Measuring population diversity. *American Sociological Review* (1969), 850–862.
15. Morris, M. R. Web on the wall: Insights from a multimodal interaction elicitation study. In *Proceedings of the 2012 ACM International Conference on Interactive Tabletops and Surfaces*, ITS '12, ACM (New York, NY, USA, 2012), 95–104.
16. O'Connell, D. L., and Dobson, A. J. General observer-agreement measures on individual subjects and groups of subjects. *Biometrics 40*, 4 (1984), pp. 973–983.
17. Scott, W. A. Reliability of content analysis: The case of nominal scale coding. *Public opinion quarterly* (1955).
18. Vatavu, R.-D., and Wobbrock, J. O. Formalizing agreement analysis for elicitation studies: New measures, significance test, and toolkit. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, ACM (New York, NY, USA, 2015), 1325–1334.
19. Wagner, J., Huot, S., and Mackay, W. Bitouch and bipad: Designing bimanual interaction for hand-held tablets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 2317–2326.
20. Wobbrock, J. O., Aung, H. H., Rothrock, B., and Myers, B. A. Maximizing the guessability of symbolic input. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '05, ACM (New York, NY, USA, 2005), 1869–1872.
21. Wobbrock, J. O., Morris, M. R., and Wilson, A. D. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, ACM (New York, NY, USA, 2009), 1083–1092.
22. Wood, M. Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods 8*, 4 (2005), 454–470.